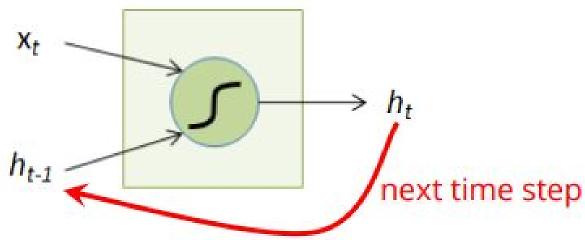
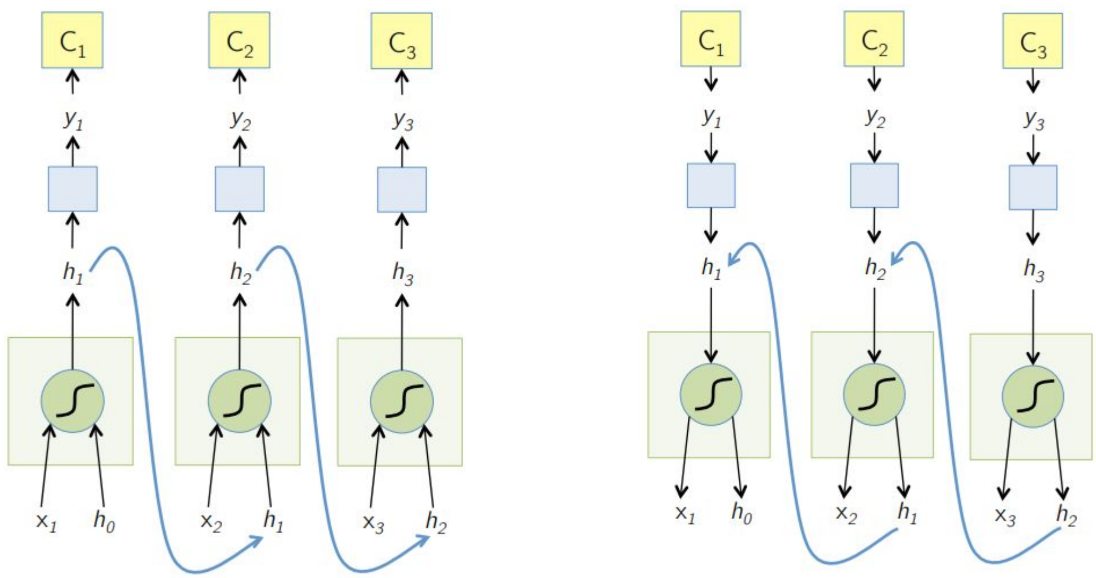
# Long Short-term Memory (LSTM)

### 0.Recurrent Neural Network



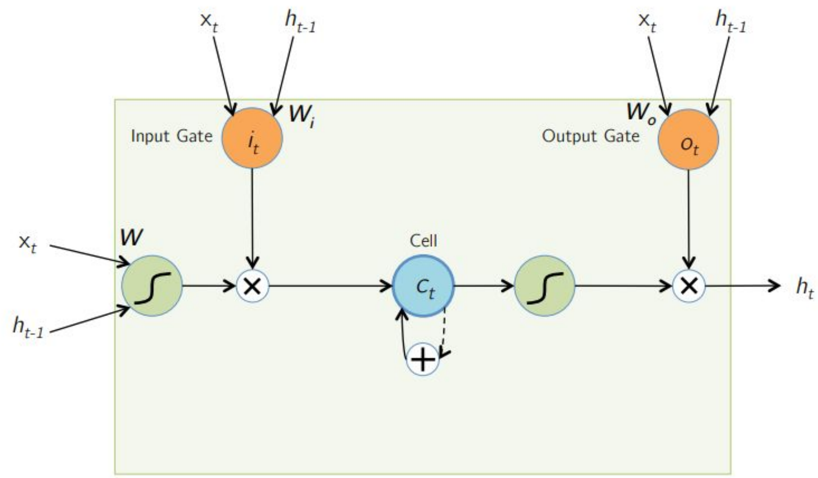

$$h_t = f(W_x x_t + W_h h_{t-1} + b)$$

### 0.1. Backpropagation Through Time  (BPTT)



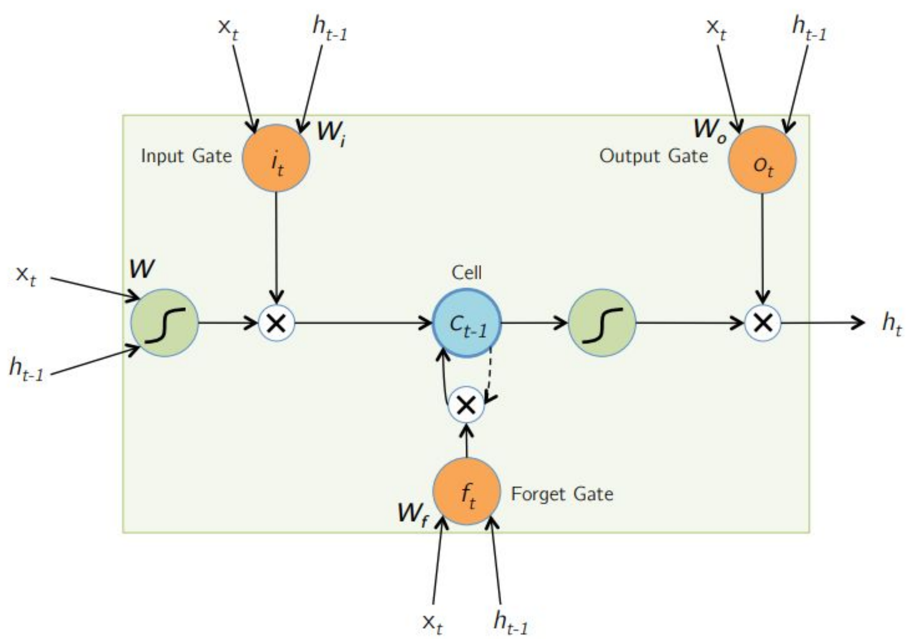

### 0.2.Real Time Recurrent Learning (RTRL online learning)

### 0.3. Constant Error Carousel (CEC)




$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$
$$a_t = tanh(W_c x_t + U_c h_{t-1} + b_c)$$
$$c_t = c_{t-1} + i_t \odot a_t$$
$$h_t = o_t \odot tanh(c_t)$$

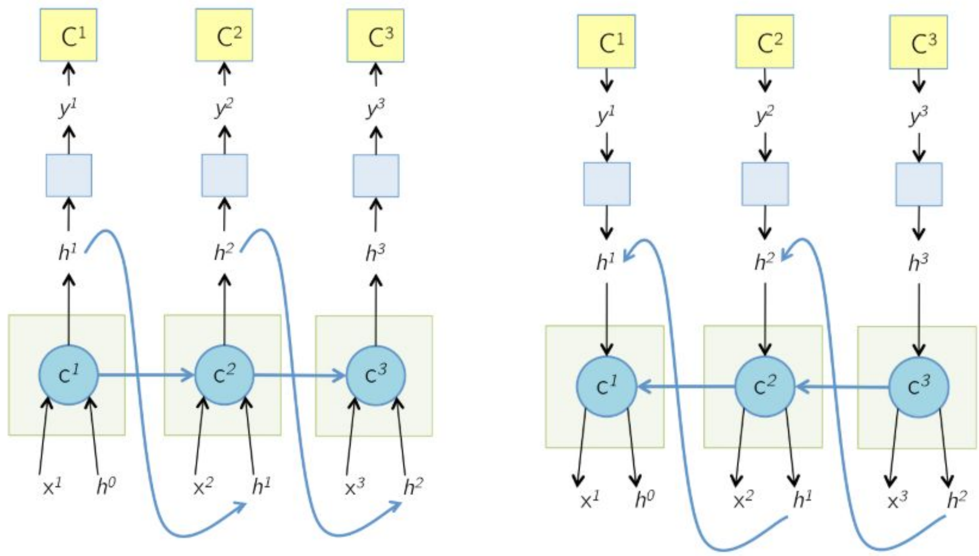
**1.LSTM**




$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$
$$a_t = tanh(W_c x_t + U_c h_{t-1} + b_c)$$

$$f_t = tanh(W_f x_t + U_f h_{t-1} + b_f)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot a_t$$
$$h_t = o_t \odot tanh(c_t)$$

**1.1.LSTM BPTT**




**1.2. Pro and Con**

**Pro:**

a. Mitigates gradient vanishing and exploding problems of rnn
b. Cell state is protected by forget gate, good for noisy sequences

**Con:**

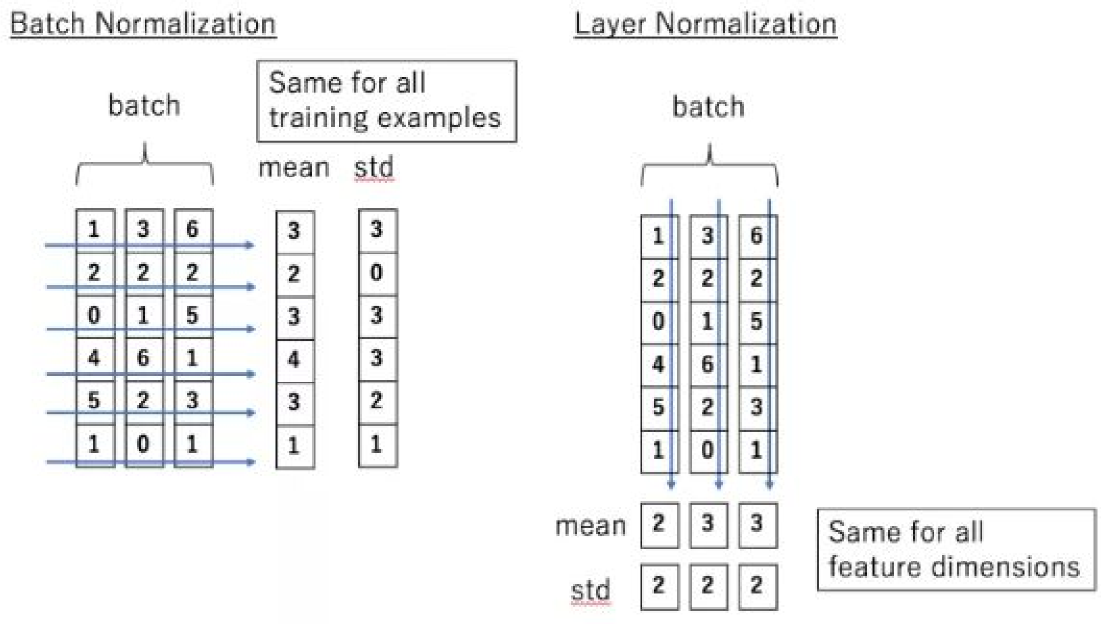
a. Long training time
b. Large memory usage
c. Overfitting

**2.Opitimize LSTM**

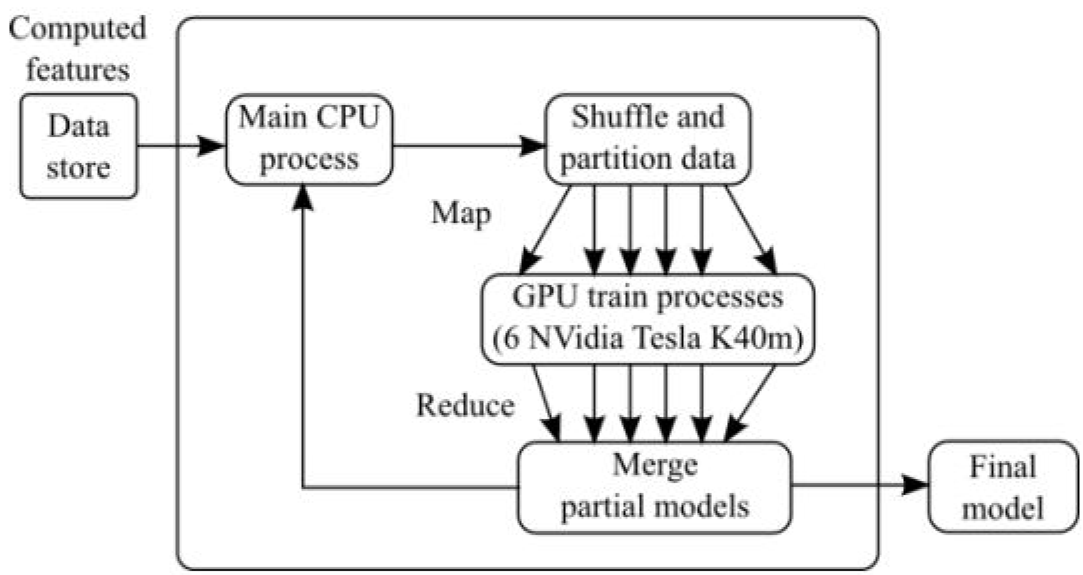
**2.1.Mini-Batch**

**2.2.Batch Normalization**

**2.3.Layer Normalization**




**2.2.GPU Acceleration**




**2.4.Truncated Backpropagation**
**2.5.Adaptive Learning Rate**
**2.6.Dropout to avoid overfitting**
**2.6.More..**

**3. LSTM Variations**
**3.1.Peephole LSTM**

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + V_i c_{t-1} + b_i)$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o c_t + b_o)$$
$$a_t = tanh(W_c x_t + U_c h_{t-1} + b_c)$$
$$f_t = tanh(W_f x_t + U_f h_{t-1} + V_f c_{t-1} + b_f)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot a_t$$
$$h_t = o_t \odot tanh(c_t)$$

### 3.2.Coupled Input and Forget Gate

$$f_t = 1 - i_t$$

### 3.3.Full Gate Recurrence

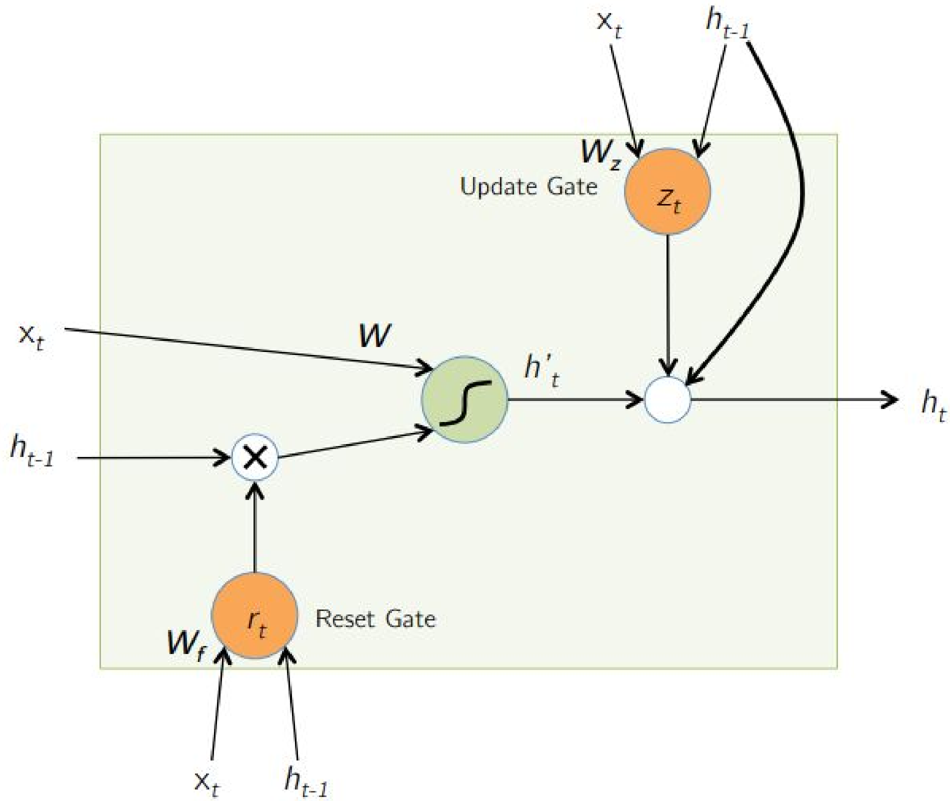
$$f_t = \sigma\left( W_f \begin{pmatrix} x_t \\ h_{t-1} \\ c_{t-1} \\ i_{t-1} \\ f_{t-1} \\ o_{t-1} \end{pmatrix} + b_f \right)$$

### 3.4. More Variants

a. No input gate $i_t = 1$
b. No forget gate $f_t = 1$
c. No output gate $o_t = 1$
d. No input activation function y=x
e. No output activation function y=x
f. No peepholes

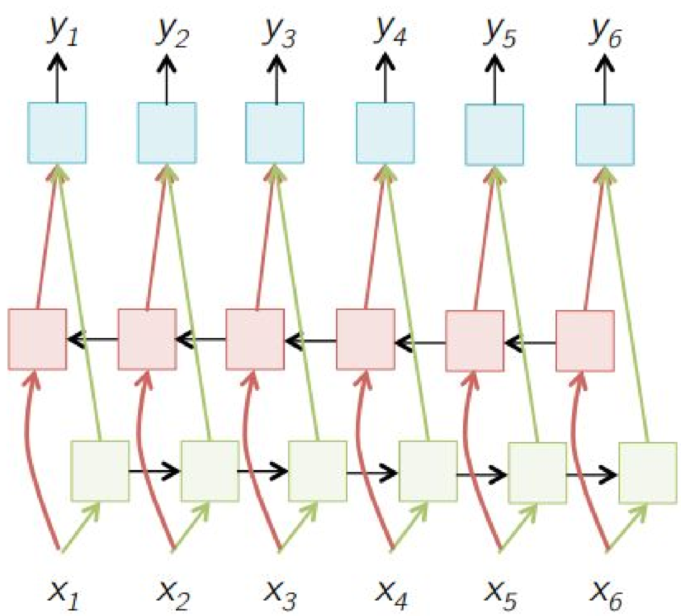
- The standard LSTM performed reasonably well on multiple datasets and none of the modifications significantly improved the performance
- Coupling gates and removing peephole connections simplified the LSTM without hurting performance much

### 3.5. Gated Recurrent Unit (GRU)

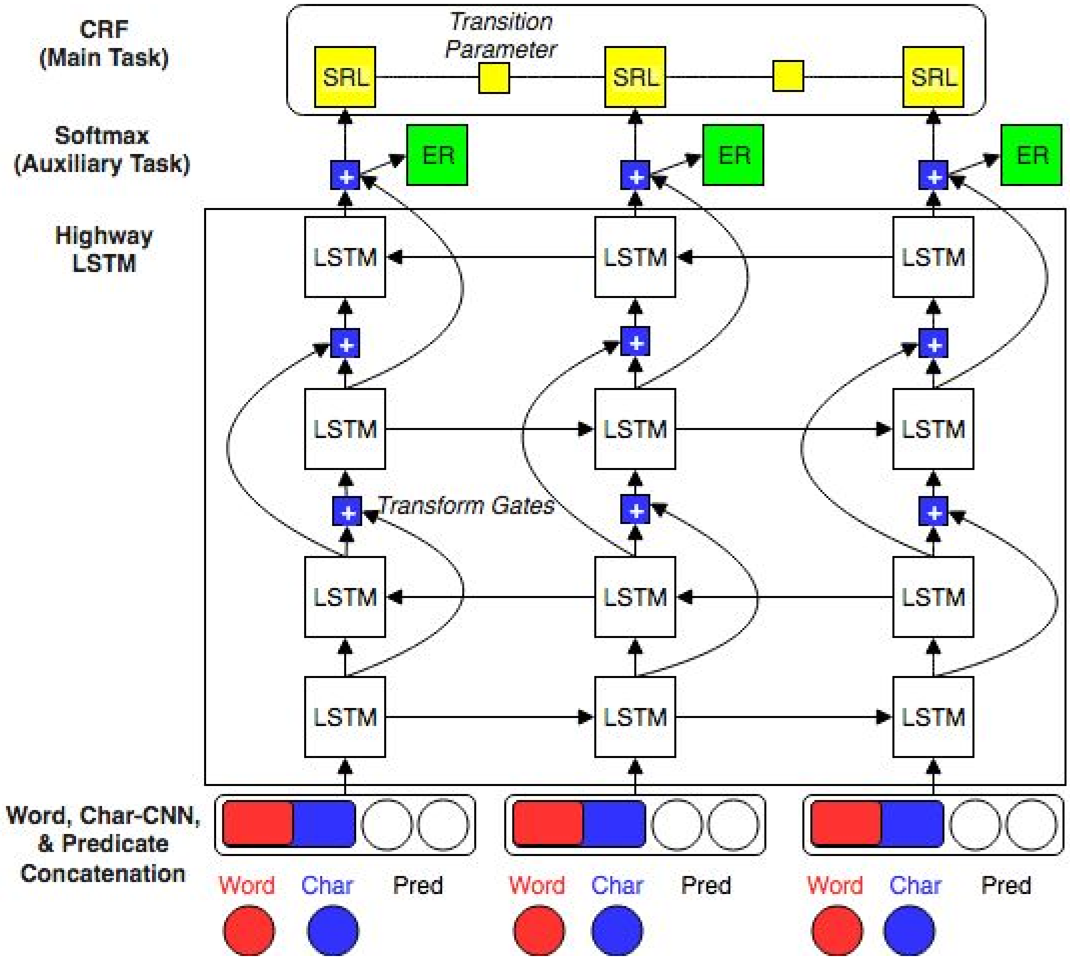



$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$
$$h'_t = tanh(W_h x_t + U_h(r_t \odot h_{t-1}))$$
$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$
$$h_t = z_t \odot h_{t-1} + (1 - z)_t \odot h'_t$$

**3.6.Bidirectional LSTM**



**3.7.Highway LSTM**




**4.Reference**

Pics source: http://slazebni.cs.illinois.edu/spring17/lec02_rnn.pdf,
http://slazebni.cs.illinois.edu/spring17/lec03_rnn.pdf
Long Short-term Memory https://www.bioinf.jku.at/publications/older/2604.pdf
LSTM: A Search Space Odyssey https://arxiv.org/pdf/1503.04069.pdf
Understanding LSTM Networks http://colah.github.io/posts/2015-08-Understanding-LSTMs/
Accelerating Recurrent Neural Network Training using Sequence Bucketing and Multi-GPU Data Parallelization https://arxiv.org/ftp/arxiv/papers/1708/1708.05604.pdf
Layer Normalization https://arxiv.org/pdf/1607.06450.pdf